CTLS

EVALUATING COMPLEX STUDENT LEARNING

P.A. Cranton

1982

Centre for Teaching and Learning Services

Evaluating Complex Student Learning

P. A. Cranton

Centre for Teaching and Learning Services

Mc Gill University

1982

Table of Contents

Complex Student Learning	5
Performance or Psychomotor Learning	5
Affective Learning	7
General Consideration in the Evaluation of Complex Student Learning	8
Planning the Evaluation Technique	9
What is Being Evaluated?	9
Selecting the Evaluation Technique	12
Evaluation of Performance	12
Checklists	13
Rating Scales	13
Indirect Measures	14
Summary	14
Evaluation of Affective Learning	15
Performance Techniques	16
Attitude Scale	16
Questionnaires	17
Discussions and interviews	17
Essays and written reports	18
Indirect measures	18
Summary	18
Constructing the Instruments	19
Direct Observation Techniques	19
Checklists	19
Rating scales	20

Indirect Measures 21
Interviews 21
Discussions 21
Anecdotal records and comments 22
Essays and written reports 23
Attitude scales and questionnaires 23
Testing and Revising the Instrument 28
Review 29
Pilot Testing 30
Rater Agreement 30
Internal consistency 31
Agreement with other measures 31
Doing the Evaluation 32
Timing 32
Observer Effect 32
Setting 33
Involvement of Others 33
Results 33
Scoring 34
Structured techniques 34
Unstructured techniques 34
Grading 36
Criterion-referenced approach 36
Norm-referenced approach 36
Alternative approaches 39
Summary 40
References 41

Evaluating Complex Student Learning

A Faculty Resource Book

The learning expected of students in many university courses is of a nature which can not easily be evaluated using the traditional multiple choice or essay test. We may expect students to develop the complex interpersonal skills required in a therapist-client relationship, to develop the ability to produce a creative work of art, or to learn the series of physical skills involved in a game of tennis. The instructor concerned with these types of learning most often faces the formidable task of developing evaluation techniques with very little guidance, and may well decide that it cannot be done. It is the purpose of this resource book to provide procedures for specifying and evaluating complex student learning.

First, the various types of student learning included under "complex learning" will be described, and some general evaluation considerations will be discussed. This will be followed by procedures for planning the evaluation, constructing the instruments, and doing the evaluation.

Complex Student Learning

Learning is most often thought of in terms of "cognitive development", the students' knowledge of definitions and terms, the demonstration of understanding by providing examples or illustrations, the application of information to solving problems, the analysis of relationships, the synthesis of ideas, and the ability to evaluate or critique others' ideas. These types of learning can nearly always be assessed on paper. Multiple choice tests and other objectively scored instruments can be used to measure basic knowledge and the application of the knowledge to specific situations. Essays, term papers, case studies, and proposals can be used to judge the students' analysis and synthesis of ideas. All of these skills are what we traditionally think of as learning, and many educators have devoted effort to the construction of appropriate tests and scoring techniques for this learning.

The remaining area of "complex" or "non-cognitive" student learning contains a diverse set of skills and abilities. For convenience, the educators' classification of this learning into two categories or "domains" will be followed.

Performance or Psychomotor Learning

This category includes all learning which is demonstrated through physical performance. Several authors have attempted to subdivide this domain into categories or taxonomies of behaviors (c.f., Dave, 1970; Harrow, 1972). The most useful sub-categories for the university instructor are probably those provided by Kibler, Barker and Miles (1970): (a) Gross body movement includes basic locomotor movements and would be a part of the fundamental skills required in many areas of physical education. (b) Finely coordinated movement includes visual motor coordination and manipulative

skills. Clinical skills, such as administering an I.V., fall into this category. The operation of equipment and instruments (in a clinic, laboratory, computer centre or media centre) may be relevant student learning in many courses. Obviously the eye-hand-foot coordination required in many sports is a part of this area of student learning. (c) Non-verbal communication is the behavior involved in portraying feelings and attitudes through body posture, movement, gestures and facial expressions. This is considered an important area of learning in interviewing, therapy, drama, and debating. A social work instructor, for example, will want to judge the students' skill at non-verbally expressing empathy toward a client. (d) Speech behaviors are also considered to be performance learning, in terms of the physical production of speech: pronunciation, expression, voice projection, etc. Instructors of second languages, drama, and speech therapy, for example, will be interested in evaluating these skills.

As can be seen, the performance or psychomotor domain of student learning includes a diverse set of behaviors and skills in many different disciplines. It should be noted that this area of learning cannot be completely separated from the cognitive domain: students usually "learn", in the traditional sense, what the steps are in performing a task or using a piece of equipment, before they learn to complete the performance. The cognitive knowledge is inseparable from the ability to perform in most cases. However, it is clear that we wish to evaluate the performance as well as the knowledge. A nursing student who can describe the insertion of a needle but who cannot actually insert it can hardly be said to have mastered this

learning objective.

Affective Learning

The second domain of student learning which will be covered in this guide is labelled "affective", and is concerned with attitudes, values, beliefs, interest and motivation. As in the previous area, or perhaps more so, this domain exists in conjunction with the cognitive domain. We are interested in teaching values or attitudes based on an understanding of the content. Interest in a subject area comes with a knowledge of that subject area. However, evaluation of learning in this domain, as before, does not come automatically with evaluation of the cognitive component.

Again, educators have attempted to delineate sub-categories of the affective domain. A brief review of one of these taxonomies (Krathwohl, Bloom, & Masia, 1964) may clarify the nature of learning in the area, and may be useful in considering the evaluation techniques in the next section of this guide. Five subcategories are defined by Krathwohl, et al. (1964). (a) Receiving is the students' awareness of and willingness to receive information. For example, they may be willing to listen to a radical theory applied to social work, or they may develop a tolerance for bizarre uses of shading in modern art. (b) Responding is the level where the students respond to affective information, e.g., they voluntarily look for instances of art in which shading has been well used, or express an interest in social problems broader than those of the local community. (c) Valuing describes an increasing internalization as the students begin to hold a value. They may, for example, develop a desire to write effectively, or actively seek out examples of good art for enjoyment, or express faith in the scientific method of experimentation. (d) Organization occurs when the learners encounter situations in which more than one value is relevant, and organize the values

into a system. The students may, for example, weigh alternative social policies and practices against the standards of public welfare. (e) Value complex is the category of student learning where the individual responds consistently to a value-laden situation with an interrelated set of values and structured view of the world. The students would, for example, be able to formulate an approach, based on values, to a particular client-therapist situation.

The affective domain of learning, as the psychomotor, is relevant in a wide variety of subject areas. Minimally, instructors in almost every area are concerned with the students' interest in the subject area. In other specialized fields (education, social work, nursing, psychiatry), the instructor is concerned with the students' development of attitudes and values which will affect the application of knowledge to practical work.

General Consideration in the Evaluation of Complex Student Learning

As can be seen from the above descriptions, student learning in the psychomotor and affective domains is not straightforward. It is clearly connected to the well-understood "cognitive" or knowledge domain, and yet it is the application of this knowledge to either performance or values or beliefs. To evaluate the latter, one must be considering actual behavior, and judging it in some way. There is a tendency to conclude that this cannot be done in a reliable way, therefore, one should concentrate on the "cognitive knowledge" of students, assuming that it will be displayed in appropriate ways. However experience has shown that the student who does well in theoretical courses does not necessarily apply this knowledge to the practical situation. Faculty in the professional schools of medicine, nursing, and social work have realized that they must evaluate their students' performance in real or simulated conditions. This evaluation must, by

definition, involve some subjective judgement as to the students' performance, or demonstration of values. It will be the task, in this guide, to first specify the behavior to be evaluated, and then to develop techniques and instruments which, as objectively as possible, assess student learning.

Planning the Evaluation Technique

What is Being Evaluated?

As in any type of evaluation, before work commences on the development of instruments, it must be clear precisely what student learning is expected. In other words, what would the student do, or say, that would demonstrate the particular performance or affective learning relevant to the course or program. Detailed guidance on writing objectives or goals is available in a number of books (cf., Mager, 1975); however, some attention will be given to this important preliminary step here.

The closer the objective, or goal statement comes to describing observable, specific or measurable student learning, the easier it will be to develop the evaluation technique. A good beginning is to write down any statement of what is expected, however vague or non-measurable it seems, then ask, "How would a student demonstrate that this had occurred?"

Example (Performance, Nonverbal Communication)

> "The student will create a warm, relaxed atmosphere in an interview situation."
>
> How might a student actually demonstrate this?
>
> - smiling appropriately
> - eye contact with interviewee (facial, and body posture)
> - appropriate distance from interviewee

Although this is not a list of precise behaviors (and need not be) it is a move toward specifying what the first statement includes. The degree of specificity desired would depend on the subject area, the level of the student and the preference of the instructor.

Example (Affective Learning, Responding)

> "The student will feel comfortable in a group discussion and significantly contribute to the conversation."
>
> How would you know when this goal was reached?
>
> - voluntarily speak in a group discussion
> - ask thought provoking questions in a discussion

Again, the "unobservables" (feel comfortable) are translated into "observables" (voluntarily speak). Further specification is still possible, for example, indicating the amount or the timing of speaking; however, it may not be necessary.

One author (Mager, 1972) refers to this process as "eliminating the 'fuzzies'." The goal at this stage is to make the statements of expected learning as clear and as observable as possible. Table 1 gives several examples of phrases and statements rewritten in a more observable format.

The next step is the selection of an evaluation technique.

Table 1

Examples of Clarified Statements

of Student Learning

Unobservable	Observable
The student likes baseball.	Reads about baseball in the sports pages. Can recite batting averages. Attends all baseball games.
The student enjoys biology.	Attends university biology club meetings. Asks questions in biology class. Reads biology books 70% of the time in the library.
The student will develop a good therapeutic technique	Recognizes patient charcteristics, to which the nurse should respond. Responds with the proper skill to patient characteristics.
The student appreciates discussion.	Allows others to finish before presenting a point of view. Accepts information presented by others.

Selecting the Evaluation Technique

Although some overlap will become apparent, evaluation methods will be discussed separately for performance and for affective learning. The overlap occurs where affective learning is demonstrated through a performance (e.g., a student who "enjoys" tennis will play frequently or will play well).

Evaluation of Performance

Evaluation of student performance or "psychomotor learning" may be of two general types: (1) process evaluation where the actual process or performance is observed and judged (e.g., observing a student perform an experiment to condition rats in a maze); or (2) product evaluation, where the performance is judged by examining the result or product of it (e.g., judging whether or not the rats have been properly conditioned). The choice between these two types of evaluation depends on variables such as: the amount of time available, the importance of the actual process, whether or not the product is likely to accurately reflect the quality of the process. Some combination of process and product evaluation may also be employed (e.g., observing a segment of the process, then judging the quality of the product).

A second choice which should be made early in the planning of the evaluation is one between real and simulated conditions. It may be important, or essential, to observe the student in a real situation, for example, actually administering nursing care to a patient. Or the real situation may yield student anxiety, or result in serious consequences if a mistake is made, and a simulated situation would be preferred (e.g., a student role-playing a therapeutic approach with a suicidal patient).

In this section of the guide, techniques will be described with emphasis on the situations in which they are most appropriate. The next section ("Constructing the Instruments") will contain guidelines for actually developing the techniques.

Checklists. Used either in a process or a product evaluation, a checklist is simply a list of behaviors (process) or characteristics (product) which should exist in, or as a result of, a given performance. The instructor and/or another observer indicates, usually while observing, whether each behavior or characteristic has or has not occurred. A checklist is used for a performance where no judgement of quality is required. For example a communications student learning to operate a television camera, or a physiotherapy student learning to assemble equipment would demonstrate a set of behaviors which either occur or not, and no degrees of quality of performance exist. Checklists may be used in real or simulated conditions, and with "live", videotaped, or audiotaped performance.

Rating scales. As with a checklist, a rating scale may be used for either process or product evaluation and under real or simulated conditions. This technique, however, incorporates a judgement of the quality of performance or the frequency of a behavior into the evaluation. It is used, therefore, when some judgement is a relevant aspect of the evaluation. A rating scale consists of a list of behaviors or characteristics which are observed along some continuum or scale of quality or frequency of occurrence. The use of colour, for example in an oil painting, may be rated from "excellent" to "poor" in a product evaluation. Or a students' use of non-verbal communication could be rated on a frequency scale, such as "always" to "never". Different types of scales and guidelines for constructing

them will be discussed in the next section.

Indirect measures. In some situations it is not possible or practical to observe or record an actual performance or product of a performance. In these cases, an instructor can turn to indirect measures such as student journals, anecdotal notes made by students or other participants in a situation (support staff, patients), records (library, computer, etc.) of student activities, or post-performance discussion with, or interviewing of individuals involved in the situation. Indirect measures must be used carefully to ensure that inferences can be made from the information to the student's actual performance; because of this inference, more "error of measurement" is introduced than in direct observation. Indirect measures are chosen on the basis of variables such as: time and personnel limitations, the interference of observation with the performance (e.g., an observer in a therapist-patient interaction could change the behavior of both individuals), or practical considerations (e.g., eight student nurses in different areas being evaluated by one instructor). In addition, they may be useful as a supplement to other techniques, revealing aspects of the performance that were unobservable. For example, a journal kept by a student teacher may reveal an incident that occurred in the classroom the day before the supervisor evaluated the performance, and which affected the student's behavior during the evaluation.

Summary. As can be seen, the measurement of affective learning is a complex undertaking. Once the goals or objectives are clearly stated, and one or more of the above techniques are selected, it is possible to assess student attitudes, values or interest. However, more caution must be used in this area than in any other domain of student learning.

At this point, selection of the technique or techniques which are most appropriate for an instructional situation should be possible:

a) Process and/or Product Evaluation

in

b) Real or Simulated Conditions

involving

c) Live, Videotaped, or Audiotaped Performace

using

d) Checklists, Rating Scales and/or Indirect Measures

Evaluation of Affective Learning

The affective domain may well be the most difficult type of student learning to evaluate, even though there has been considerable work done in the area of attitude measurement. First, interest, motivation and even attitudes are often difficult to translate into specific or observable statements. Second, the link between affective learning and performance may be a weak one; i.e., does the student perform in a certain manner because the instructor is observing or because the instructor is looking for the demonstration of a particular attitude. Third, if the affective learning is assessed through rating scales, questionnaires, or any written product, the student may respond according to perceived expectations.

Evaluation techniques for the affective domain will be described below, and the situations in which they can be used will be discussed. Given the above cautions, it may be advisable to use more than one technique, particularly if the affective learning comprises a major component of a

course or program.

Performance techniques. As pointed out at the beginning of this section, affective learning may be demonstrated through student performance. When this is the case, the techniques discussed for evaluating student performance may be employed. For example, if a student of special education is expected to develop an "attitude of respect" toward the handicapped child, this might be specified in terms of a list of nonverbal and verbal communications to be observed in an interaction between the student and the child. A rating scale could then be employed to examine the frequency with which these behaviors occur. The reader is referred back to the previous section for a review of the techniques of performance evaluation.

Attitude scale. The most common and well-researched technique for evaluation in the affective domain is the attitude scale. Many standardized, published instruments are available (e.g., College Interest Inventory, Differential Value Profile); these are described and reviewed by Buros (1972) in the Seventh Mental Measurements Yearbook. In addition, many books and articles discuss the development of attitude scales and the techniques for refinement (Payne, 1974; Henerson, Morris, & Fitz-Gibbon, 1978). If attitude scales are used extensively, it would be useful to refer to the more detailed information. The technique will be described briefly in this section of the guide, and some general guidelines for constructing scales will be discussed in the next section.

Although an attitude scale can have a variety of formats, it generally consists of a series of statements expressing opinions on one or more topics, and asking students to rate or rank the degree to which the statements express their own opinions. Responses are recorded as numbers and summed

to yield an attitude score on one scale or on more than one subscale. A social work professor, for example, may be interested in the value that students place on research in relation to their own practice. Students could be given statements such as "Social work practice should be modified on the basis of research findings" and asked to respond to the statements on a "Strongly Agree" to "Strongly Disagree" scale.

An attitude scale is appropriate when an instructor wishes to quickly assess opinions from a large number of students, and when there is little reason for the students to "fake" the responses (e.g., if the instructor has not expressed an attitude, or has expressed acceptance of a variety of attitudes).

Questionnaires. Questionnaires present information to the student in the form of a statement or question and can ask for a variety of responses, from a rating to a phrase or several sentences. An open-ended format allows the student flexibility in answering and may uncover attitudes or values that the instructor had not foreseen. Similar to the attitude scale, a questionnaire can be used to efficiently collect a variety of information from a large number of students.

Discussions and interviews. Depending on the size of the class and the nature of the affective learning, dialogue with the students can assist in the evaluation. Structured discussions or interviews, conducted in a group or individually will allow the instructor to ask a series of questions concerning attitudes, values, or interests. It might, for example, be a goal of an instructor of medical ethics to teach students a tolerance of values different from their own. In an interview, the instructor might ask, "What would be your response to a 15 year old girl requesting an abortion?" More informal class discussions can provide an instructor with information about student

interest or motivation. It may even be useful to regularly record comments made during such discussions as the course progresses.

Essays and written reports. Although essays, terms papers, case studies, and reports are usually viewed as vehicles for the assessment of cognitive learning, they can also be used in the evaluation of affective learning. A counselling instructor, for example, may want students to be aware of and accept a variety of societal values. An essay or case study involving a controversial issue could be used to reveal such an aspect of development. Students' areas of interest could also be explored by asking students to write on an issue of their choice related to course content. At times an instructor may wish to do a systematic analysis of students' attitudes or values, using the written essay or paper; this could lead to a "content analysis", which will be referred to in a later section of the guide.

Indirect measures. When used cautiously, indirect measures of student interest or attitude can yield useful evaluation information. Library records of student use of reserved books or articles, could, for example, be one indication of interest in a particular topic. A follow-up on the number of students taking further courses in an area reveals interest in the subject area. Class attendance, informal observation of student activities (do the students in your tennis class remain to play tennis after the instruction?) and anecdotal notes or records of class events can provide supplemental information as to student interest or motivation.

Summary. As can be seen, the measurement of affective learning is a complex undertaking. Once the goals or objectives are clearly stated, and one or more of the above techniques selected, it is possible to assess student attitudes, values or interest. However, more caution must be used in this

area than in any other domain of student learning.

Constructing the Instruments

Guidelines for constructing each of the instruments referred to in the previous section will be presented. It may be the case that an instructor is able to use or adapt an instrument which has already been developed; Colleagues at other universities and professional associations are often useful sources of evaluation techniques.

Direct Observation Techniques

When goals or objectives are clearly written, the development of a checklist or rating scale becomes relatively straightforward.

Checklists. A checklist is used to determine whether or not a set of behaviors or characteristics has occurred, and should consist of a clear, specific list of items. General guidelines include:

1. Items should be short, clear, and unambiguous--different observers should interpret each item in the same way.
2. Items should contain only one behavior or characteristic.
3. Items should be objective, answerable in a "yes-no", "exist-doesn't-exist" format.
4. The number of items should be kept to a minimum--some authors recommend less than ten.
5. The items should be organized in a logical sequence, or in categories to facilitate the observer's use of the checklist.

In any observational evaluation, the validity and reliability of the observation become important--if necessary train the observer or allow a "practice" session. At times, it may be important to use more than one observer and to check the consistency of their recordings. However, generally, the use

of checklists, when developed according to the above guidelines, is a valid and reliable technique.

Rating scales. The judgement of the quality of performance introduces more subjectivity into the evaluation and each of the issues raised previously become even more relevant. General guidelines for the construction of a rating scale include:

1. Items should be clear, objective, and unambiguous--two raters should agree as closely as possible on the meaning of the item.
2. Each item should contain only one behavior or characteristic.
3. All points on any scale used should be clearly defined as to the degree of quality or the frequency they indicate.
4. An equal number of positive and negative points should occur on a scale, and there should be equal intervals between points on a scale.
5. A seven-point scale is the maximum number that individuals have been shown to reliably respond to. A four- or five-point scale is most commonly used.
6. Items should be placed in a logical sequence or grouping to facilitate location of the item for the observer (e.g., order of occurence, similar types of behaviors together).
7. Formatting is important--items should not be forced on to a page, spacing should be used to facilitate the location of an item.

In the construction of a rating scale, developmental testing of the instrument is important, unless it has already been used and tested under similar circumstances. Colleagues and/or students can be asked to comment on clarity, comprehensiveness, and ease of use. Observers may need to be trained, and the consistency of ratings made by two observers should be checked, when possible, before the instrument is actually used to evaluate

student learning.

Indirect Measures

The use of indirect measures may not require advance preparation of instruments. The use of records, for example, (library, computer, registration) requires only that there are clearly specified goals for assessing and interpreting the information. However, the use of interviews, discussions, anecdotal records, and comments do require careful advance consideration.

Interviews. Interviews used to evaluate student performance or affective learning shoud be well-planned and should consist of a series of structured or semi-structured questions relating directly to the learning that is expected to take place. General guidelines for formulating interview questions include:

1. Be specific and direct--ask what you want to know, leaving no room for interpretation.
2. Be clear--avoid jargon or vague phrases.
3. Be concise--students may be reluctant to commit time to the endeavor.
4. Be flexible--useful information can be obtained if people have an opportunity to express themselves freely at some point.
5. Be objective--your questions should not lead the student to the "appropriate" answer.
6. If someone else is conducting interviews, be sure that they understand the questions being asked, and the purpose of the interview; if necessary provide training or practice.
7. If possible, plan to tape-record the interviews, or develop a clear form to record responses during the interview.

Discussions. Discussions will tend to be less structured and more subjective than interviews. As such, they are probably most valuable in assessing values, attitudes, and interest. As with interviews, it is valuable

to pre-plan the questions to be asked and to provide as much structure as possible while remaining flexible and encouraging students to express themselves. The guidelines presented for the use of interviews also apply to discussions. In addition, some of the following suggestions may prove useful:

1. Circulate in advance a copy of the questions to be discussed.
2. During discussion, use time limits as a means of encouraging "to the point" discussion.
3. Request examples to yield more specific information.
4. Periodically summarize the major points raised in the discussion.

Anecdotal records and comments. If careful planning goes into the collection of anectodal notes and comments, they can provide valuable information. The same "structured-yet-flexible" technique applies here. Some suggestions for generating useful records and comments are:

1. Determine in advance what behaviors will be recorded.
2. Record only the "what" or "how"; avoid subjective interpretation of the event (unless students are recording their own reactions to an event).
3. Describe the context in which records are made or comments collected.
4. Collect comments or records as soon after the incident or observation as possible.
5. Collect as much information as possible on any one student before drawing conclusions ("repeated measures" will increase the reliability of the information).
6. Be flexible--accept unsolicited comments that are relevant to your evaluation.

Essays and written reports. Essays, reports and papers are often useful in the assessment of affective learning. In general, their value depends on the topics given or questions asked, the guidelines given to students and the care taken in analyzing the written work. Some suggestions for presenting this task to students include:

1. Provide clear and complete directions, including the purpose or purposes of the essay or report.
2. Be sure that the topic stimulates interest and provides the leeway for the expression of opinions.
3. Encourage students to express their own feelings rather than what they have read or what they think is expected of them.
4. The wording of the question or topic should not suggest the attitude to be taken.
5. Be sure that the time allowed for responding is adequate; students under a time pressure will not write with as much care.

In order to analyze students' writing for the measurement of affect, consider the following guidelines:

1. Pre-determine areas or categories into which responses will be divided (e.g., a positive attitude toward euthanasia).
2. Analyze only the content relevant to the attitude(s) or value(s) in which you are interested.
3. Be as objective as possible; check your analysis with a colleague, or re-do some part of it at another time and check for consistency.
4. Quantify the results, if possible; frequency counts and percentages can be very helpful.

Attitude scales and questionnaires. The development of a reliable and valid attitude scale is a difficult and time consuming procedure. Whenever possible, it is preferable to use or adapt items which have been written, tested, and/or published elsewhere. Some of the difficulties encountered in constructing your own scale will be discussed briefly.

Assessing attitudes or values with a paper and pencil inventory is a somewhat artificial procedure. The relationship between how a student responds on a scale, and the actual behavior may be questioned. A student is often able to determine what type of information the professor hopes to receive (e.g., the value of research in social work practice) and to answer accordingly. One particular type of distortion on which research has been conducted is called "social desirabilty": the attempt of the examinee to present a socially acceptable image. In addition, psychologists and educators have found that a "response set" often exists. That is, some individuals, when unsure of a response, will tend to choose the "agree" category or the "neutral" category, regardless of the content of the item.

In spite of these difficulties, situations will arise where it is useful to construct an attitude scale. If caution is used, students' affective learning can be assessed with this technique.

Edwards (1957) provides guidelines for developing statements to be used in attitude scales.

1. Avoid statements that refer to the past rather than to the present.
2. Avoid statements that are factual or capable of being interpreted as factual.
3. Avoid statements that may be interpreted in more than one way.
4. Avoid statements that are irrelevant to the psychological object under consideration.
5. Avoid statements that are likely to be endorsed by almost everyone or by almost no one.
6. Select statements that are believed to cover the entire range of the affective scale of interest.
7. Keep the language of the statements simple, clear, and direct.
8. Statements should be short, rarely exceeding 20 words.
9. Each statement should contain only one complete thought.

10. Statements containing universals such as all, always, none, and never often introduce ambiguity and should be avoided.

11. Words such as only, just, merely, and others of a similar nature should be used with care and moderation in writing statements.

12. Whenever possible, statements should be in the form of simple sentences rather than in the form of compound or complex sentences.

13. Avoid the use of words that may not be understood by those who are to be given the completed scale.

14. Avoid the use of double negatives.

Several scale construction techniques have been suggested by measurement experts. A few of the most common and least difficult to use methods will be discussed.

Forced choice selection method. A forced choice item requires the respondent to select one of two or more statements that differ in content. Students are asked to indicate which action or charactertic is most applicable to them. Two example items taken from the Preferred Job Characteristics Scale are:

I prefer:

1. A job where my opinion is valued.

or

2. A job with short working hours.

I prefer:

1. A job which does not tie me down.

or

2. A job where I could decide how the work is to be done.

A multiple choice format (where more than two alternatives are provided)

may also be used. For example:

> Viewing Leonardo da Vinci's picture, "The Last Supper", would you tend to think of it:
>
> a) as expressing the highest spiritual aspirations and emotions.
>
> b) as one of the most priceless and irreplaceable pictures ever painted.
>
> c) in relation to Leonardo's versatility and its place in history.
>
> d) as the quintessence of harmony and design.

Likert scales. A Likert scale contains a set of statements representing both positive and negative attitudes. Students are then asked to select one point on a scale which indicates their agreement with the statement. The usual response categories are: strongly agree, agree, neutral, disagree, and strongly disagree. Weights of 1 through 5 are assigned to the points on the scale so as to yield a higher numerical score for a more positive attitude. Example statements for a scale measuring attitude toward education could include:

1. I am intensely interested in education.
2. Education does far more good than harm.
3. If anything, I must admit a slight dislike for education.
4. I go to school only because I am compelled to do so.

Students would be asked to indicate the extent to which they agreed or disagreed with each statement.

Semantic differential technique. Developed by Osgood, Suci, and Tannenbaum (1975), the semantic differential provides an efficient and well-researchd technique for measuring attitudes toward concepts. Students

are presented with a concept and a set of bipolar adjectives which could be used to describe that concept. Students respond on a seven-point scale, between each pair of adjectives. An example of a semantic differential is given below.

HOMOSEXUALITY

Good	___:	___:	___:	___:	___:	___:	___:	Bad
Tasteful	___:	___:	___:	___:	___:	___:	___:	Distasteful
Brave	___:	___:	___:	___:	___:	___:	___:	Cowardly
Clean	___:	___:	___:	___:	___:	___:	___:	Dirty
Happy	___:	___:	___:	___:	___:	___:	___:	Sad
Healthy	___:	___:	___:	___:	___:	___:	___:	Sick

Ordinarily about 10 to 15 adjective pairs are sufficient to measure one concept. The developers of this technique provide lists of adjectives which have been found reliable in research.

Questionnaire. A less structured technique for collecting information on affective learning, the questionnaire allows open-ended questions in addition to scaled responses or checklist items. The same meaurement problems are encountered as in the use and construction of attitude scales. Consequently, caution must be exercised in the writing of items or questions.

Bledsoe (1972) provides six criteria of a good questionnaire:

1. Brevity;
2. Inclusion of items of sufficient interest to attract the attention of the respondent and cause him to become involved in the task;
3. Provision for depth of response in order to avoid superficial replies;
4. Wording of questions neither too suggestive nor too stimulating;
5. Phrasing of questions in such a way as to alloy suspicion about hidden purposes and not to embarrass or threaten the respondent; and
6. Phrasing of questions so that they are not too narrow in scope, allowing

the respondent reasonable latitude in responses.

Although there is a slight overlap, Lauvin's (1969) suggestions may be especially relevant for instructors planning classroom use.

1. Try to limit the length of questionnaire. If the student becomes impatient to finish, he is not likely to consider the answers carefully.
2. Make sure students understand the purpose of the questionnaire and are convinced of the importance of responding completely and candidly.
3. If possible use a logical sequence of questions; this may stimulate attitudinal responses.
4. Control the administration of the questionnaire to prevent students from talking with one another about the questions before answering them.
5. Urge the students to express their own thoughts, not the responses they think the instructor wants.
6. Be sure the directions are clear, definite and complete.
7. Urge students to ask about questions that are unclear to them.
8. If possible, try out the questionnaire with one or two students to identify ambiguous questions, difficult terms or unclear meanings.

Testing and Revising the Instrument

Several times in the preceding pages, testing of an instrument or technique was referred to. In the evaluation of performance or affect, unlike cognitive learning, the judgement of an observer or rater, or the students' interpretation of questions is the basis of the assessment. It becomes, therefore, particularly important to assess the reliability and validity of the instruments being used. Would two observers watching the same student behavior rate it or categorize it in the same way? Would two

students reading a question interpret it in the same way? If not, it may well be that the information obtained is strongly affected by flaws in the instrument and is misleading. Although the "ideal" instruments probably cannot be obtained, the more obvious distortions can be eliminated with relatively little effort. Some methods for checking the quality of an evaluation technique will be discussed.

Review

Ambiguities, "leading" questions, the use of jargon, etc. can most often be noted by having someone review the instrument who has not been involved in its development. Generally, a reviewer should be a colleague, student, or graduate student who is familiar with the subject area, although in some cases, a totally naive reviewer can also be useful. Guidelines for the review process are listed below.

1. Try to have more than one reviewer, and if at all possible, include reviewers of different types (e.g., one colleague, one current student, one graduated student).
2. Give each reviewer a list of specific criteria to consider. This list may be adapted from the guidelines given in this document for the development of the instrument. Avoid merely asking for "comments".
3. Provide the reviewer with the goals or objectives that the instrument is assessing.
4. Give the reviewers feedback--let them see how their comments have been incorporated with the revised instrument, or indicate why they have not been used.
5. Encourage reviewers to also include unsolicited comments on any other aspect of the instrument.

Pilot-Testing

Pilot-testing, or an actual small scale try-out of the instrument, will likely follow the review process if both techniques are used. Pilot-testing involves using the instrument on a small number of students who are as similar as possible (e.g., level, age, experience) to the students who will be assessed using the revised version. This is an extremely valuable procedure, and a variety of types of information can be obtained.

Generally, the final evaluative situation should be replicated as closely as possible in the pilot-testing: use the same directions, similar students, and a similar environment. Do not, for example, conduct a pilot-test of a performance test in a simulated environment if the instrument will be used in a real environment.

Different types of information which can be obtained from pilot-testing will be discussed. Depending on the nature of the evaluation one or more of these categories will likely be relevant.

Rater agreement. Pilot-testing will give you an opportunity to check for agreement between raters. If the instrument is reliable, two raters should show a similar pattern of ratings or should select similar items on a checklist (a correlation of about +.70 or higher should be obtained between two raters). Where agreement is not obtained, discuss the item with the individuals involved, and try to determine the source of discrepancy. Revise the wording of the item so that raters agree on its interpretation. In some cases, it may be necessary to break an item into two or more specific parts,

or even to discard it completely.

Internal consistency. A group of items on a rating scale, attitude scale, or questionnaire which are intended to measure the same goal or objective should be responded to in a consistent manner. This may be checked by calculating correlations among those items, or if the answers are not numerical, or the pilot group is very small, by reviewing the answers carefully. If such groups of items are not answered consistently, they are likely not measuring what you intend them to measure, or they are being misinterpreted by students or raters. Discussion of the items with the pilot group will often point out the source of the difficulty, and revisions can be made in the content or wording of the items. It may be necessary to return to the review process, asking reviewers to specifically consider the internal consistency of a group of items.

Agreement with other measures. If any other measures of the same learning are available for your pilot group, check the agreement between the instrument being developed and the second measure. It might, for example, be possible to relate the results of an attitude scale pilot-testing with results of an interview or with an observation or report on actual behavior. If it is possible to calculate a correlation, the expected value would be about +.4 or +.5 or above. Again, discuss discrepancies with the students involved. It is important not to automatically assume that the fault lies in the instrument under development; several other variables could affect results, such as a lapse in time, or difficulties with the second measure.

Doing the Evaluation

Even when the utmost care is taken in the construction of an instrument, its testing and revision, problems can occur in the actual collection of information which may distort the results. An awareness of these possible difficulties will usually minimize their effect. Variables which should be considered in the evaluation of complex student learning will be discussed.

Timing

In any area of student learning, the timing of the evaluation is relevant, but this may be particularly true for performance or affective learning. Research has shown that frequent, regular feedback increases student learning. Rather than assessing student performance once at the end of a course or program, consider short frequent evaluation sessions, minimizing student anxiety and providing the feedback essential for improvement.

Observer Effect

Evaluations which are done through direct observation always run the risk of student performance being affected by the prescence of the observer. The use of one-way mirrors or even videotape may reduce this effect. If it is suspected that students are performing differently when they are not observed, have them complete a self-evaluation using the same instrument. Large discrepancies between the self-evaluations and the observer evaluations might be indicative of a difficulty in this area.

Setting

It is sometimes the case that students are taught a particular set of skills in a simulated or role-playing situation, and are evaluated in a real setting, or vice versa. An unfamiliar environment or changes in environment will affect student performance. It is important that students be familiar with the setting in which they are to be evaluated, and if possible, be allowed the opportunity to practice in that setting.

Involvement of Others

Evaluations in the clinical setting or in any situation involving other persons (patients, clients, support staff, pupils) can be affected by the behavior of the other persons. A particularly unco-operative patient, for example, will have an obvious effect on a nursing student who is being evaluated on her ability to interact with patients. In such "other person" situations, students may be asked to make verbal or written comments on the difficulty they encountered in the situation; these comments should then be considered in any judgements made. Also, it is valuable to evaluate the student in more than one situation. Repeated measures over time will usually allow the evaluator to determine the effect of the patient, client, etc. on student performance.

Results

For most of the techniques discussed in this guide, if the instrument has been carefully constructed, scoring is a relatively straightforward procedure. The exceptions to this (e.g., interviews, discussions, anecdotal records) will be discussed briefly. However, evaluation goes beyond scoring. The word itself includes an aspect of overall judgement. There is often a need to rank students, to grade them, or at least to discriminate between

those who have passed and failed. This is more likely to be the case with psychomotor than with affective learning. The second part of this section will deal with some relevant aspects of grading.

Scoring

Structured techniques. When the more structured evaluation techniques such as checklists, rating scales, attitude scales, some questionnaires, and some records are used, scoring is usually a matter of obtaining frequencies, sums or averages of ratings, or percentages of responses in a category. Scoring is complicated when, however, not all items are considered to be equally important, and it is necessary to assign weights or different scores to different items. There is no one technique for determining weights or scores for items; commonly one or more of the following are considered: (a) the amount of instructional time spent on the objective which the item or group of items is measuring, (b) the difficulty or complexity of the learning, (c) the degree to which the learning is essential in further courses, or in practice, and (d) expert opinion on the importance of the learning. Weighting systems in some subject areas (e.g., medicine) have been carefully devised, and may be valuable as a guide.

Unstructured techniques. The instructor faced with tape recordings of interviews, or a stack of essays or student journals has a much more formidable task. The formulation of well-specified learning goals, as discussed in an early section of this guide, is a necessary prerequisite to systematic, objective scoring of any of the unstructured instruments. The goals or objectives are then used to establish categories or types of information being looked for. The analysis may consist of counting the frequencies of responses in each category, of sorting students into categories dependent on their responses, or of rating students in one or more

pre-determined areas. Anyone who has attempted the analysis of verbal or written material will realize the extent to which this simplication overlooks the difficult judgements involved. Some general guidelines will be given, and other resources referred to for greater detail.

1. For each question asked in an interview, or for each goal (in the review of written material) determine the categories of responses that you would expect. It may be necessary to review a sample of your material in order to determine these categories.
2. Alternatively, all students responses or materials may be reviewed and sorted according to similarity on certain dimensions. Once there are several "stacks", determine what category each falls into.
3. For either of the above two methods, ask a colleague, if possible, to independently generate categories of responses or sorted piles and check for agreement with the first results. Discuss the discrepancies and attempt to eliminate any large differences.
4. In analyzing the results be aware of the common errors: (a) omission of some fact or relevant piece of information, (b) addition (adding to or exaggerating the respondent's reply), (c) substitution of words or phrases which have a different connotation, and (d) transposition (changing the sequence of events or overlooking the relationships among responses).
5. Be sure to consider any extenuating circumstances such as interruptions, timing, student anxiety or an unrepresentative sample of students.

If the analysis of unstructured information plays a large part in the evaluation and/or if major decisions will be based on the results, the more detailed writings of, for example, Payne (1974) or Henerson et al. (1978) may be helpful.

Grading

Grading procedures for complex student learning will not be intrinsically different from grading procedures for cognitive learning. First, two general approaches (almost "philosophies") will be discussed, followed by some alternatives to traditional grading which may be particularly relevant to affective or psychomotor learning.

Criterion-referenced approach. It is often the case, particularly for complex learning, that there is no need to compare or rank students. The priority may be that each individual student meets some criteria of performance. For example, if a physiotherapy student is learning the technique of moving a patient from a wheelchair to a set of parallel bars, the instructor may only wish to judge whether that task is successfully accomplished. This leads to a criterion-referenced approach--each student's performance is compared only to a criterion or criteria and no attempt is made to compare students to each other.

Norm-referenced approach. Traditionally grading has followed a norm-referenced approach where students are evaluated relative to their peers' performance. The well-known "grading on a curve" technique falls into this category. Any standardized test which provides "norms" or averages against which students are compared is also norm-referenced. Table 2 compares the major characteristics of norm- and criterion-referenced evaluation. In the grading of complex learning, it may be the case (e.g., in the health professions) that students must be evaluated in terms of standards established by a professional association. Instructors who, for example, are involved with the preparation of certification examinations will be using a norm-referenced approach.

Table 2

Norm-Referenced Versus

Criterion-Referenced Testing

Norm-referenced	Criterion-referenced
Reference points are average, relative points.	Reference points are fixed at specified cut-off points.
An individual's performance relative to group performance is evaluated.	Individual performance in comparison to a fixed standard is judged.
A student is evaluated as "below grade level" or "above grade level."	Grade level descriptions are not relevant.
The test fails to indicate which individuals have mastered the spectrum of instructional objectives.	Individuals who have mastered the spectrum of instructional objectives are identified.
The evaluation is vague in relation to the instructional content.	The evaluation is content-specific.
The evaluation is more summative than formative.	The evaluation is more formative than summative.
Mastery and/or success is not operationally defined.	Mastery and/or success is operationally defined.
The approach is not relevant to individualization of instruction.	The approach applies directly to the individualization of instruction.
Standardized tests are classical examples.	Tests do not tend to be standardized.

* "Summative" refers to end of unit (course, etc.) decision-making; "formative" refers to constant revisions aimed at improvement <u>while</u> instruction is taking place.

Table 2
(continued)

Norm-Referenced	Criterion-Referenced
The instrument is not sensitive to the effects of instruction.	The instrument is very sensitive to the effects of instruction.
There is a low degree of overlap with actual objectives of instruction.	Evaluation is directly referenced to the objectives of instruction.
Test items are evaluated in reference to persons.	Test items are evaluated in reference to instructional objectives.
Test results are interpreted in reference to a person's position in relation to the scores of others.	Test results are interpreted in reference to a person's position in relation to the curriculum.

From: Housden, J., & LeGear, L. An emerging model: criterion-referenced evaluation. Thrust for Educational Leadership, 1973, 2, 41-46.

Alternative approaches. Although almost any approach to grading could be classified as norm- or criterion-referenced, or some combination of the two, there are some specific alternatives which may be of interest to the instructor who is evaluating complex learning.

1. Contracting. The use of contracts for grading could be considered a criterion-referenced approach and may be particularly applicable to psychomotor learning. It is agreed in advance between the instructor and student(s) that if the student does a certain type, quantity or quality of work, a pre-specified grade will be assigned. Considerable advance planning is required and careful student files must be maintained; however, the prospect of completing the mark submissions becomes less formidable.

2. Self-Grading. In some areas of student learning (e.g., the affective domain) it is useful to have students determine their own grades. Criteria need to be determined in advance and provided for student use. The instructor may wish to supplement or comment on the student self-evaluations; this can provide a valuable learning experience for the student in the process of evaluation. In spite of reservations to the contrary, most instructors find that students are fair and objective in a self-evaluation, and often judge themselves harder than others would do.

3. Growth Grading. At times, an instructor may wish to give priority to the amount of change or growth that a student has demonstrated rather than the achievement of an absolute criterion, or a ranking relative to peers. This may be true in the case of attitudes, beliefs, or values, or in some areas of physical education (the student who increases from 0 to 20 sit-ups being as highly evaluated as the student who moves

from 20 to 40). Again, careful prior consideration must be given to the type and amount of growth and the criteria upon which grades will be based.

Summary

This resource book discussed the various types of non-cognitive or complex student learning, and procedures for its evaluation. The evaluation process, in general, consists of the following steps: (1) determining the type and level of learning which you wish your students to achieve, (2) specifying that learning in as observable terms as possbile, (3) selecting a practical and appropriate evaluation technique, (4) selecting, adapting and/or developing instruments, (5) reviewing and/or pilot testing "home-made" instruments, and (6) administering the instruments, scoring and grading students if appropriate.

References

Bledsoe, J.C. Essentials of educational research. 2nd Edition. Athens, Georgia: Optima House, 1972.

Buros, O. K. The seventh mental measurements yearbook. Highland Park, New Jersey: The Gryphon Press, 1972.

Dave, R.H. Psychomotor levels. In B.J. Armstrong et al. (Eds.) Developing and writing behavioral objectives. Tucson, Arizona: Educational Innovations Press, 1970.

Edwards, A.L. Techniques of attitude scale construction. New York: Appleton-Century-Crofts, 1957.

Harrow, A. A taxonomy of the psychomotor domain. New York: David McKay Co., 1972.

Henerson, M., Morris, L. & Fitzgibbon, C. How to measure attitudes. Beverley Hills: Sage, 1978.

Housden, J. & LeGear, L. An emerging model: criterion-referenced evaluation. Thrust for Educational Leadership, 1973, 2, 41-46.

Kibler, R.J., Barker, L.L., & Miles, D.T. Behavioral objectives and instruction. Boston: Allyn and Bacon Co., 1970.

Krathwohl, D.R., Bloom, B.S., & Masic, B.B. Taxonomy of educational objectives. Handbook II: The affective domain. New York: David McKay Co., 1964.

Mager, R.F. Goal analysis. Belmont, Calif.: Fearon Publishers, 1972.

Mager, R. F. Preparing instructional objectives, Belmont, California: Fearson Publishers, 1975.

Osgood, C. E., Suci, C. J. & Tannenbaum, P. H. The measurement of meaning. Urbana, Ill: University of Illinois Press, 1975.

Payne D. A. The assessment of learning: cognitive and affective. Lexington, Mass: D. C. Heath & Co., 1974.

Sawin, E.J. Evaluation and the work of the teacher. Belmont, California: Wadsworth, 1969.